DOCUMENT RESUME

ED 127 382 UD 016 191

AUTHOR Green, Robert L.; And Others
TITLE Standardized Achievement Testing: Some Implications for the Lives of Children.
PUB DATE Dec 75
NOTE 48p.; Paper prepared for the National Institute of Education Test Bias Conference (Washington, D.C., December 2-5, 1975)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Achievement Tests; Curriculum Development; Educational Opportunities; Elementary Education; Ethnic Groups; *Low Income; *Minority Groups; Negro Youth; Racial Differences; *Social Differences; Spanish Speaking; Standardized Tests; *Test Bias; Testing Problems; Test Interpretation

ABSTRACT

Black, Puerto Rican, Chicano, native American and low income white children represent the vast educational underclass who are most likely to be affected by test misuse or abuse. More than 50 million American children take at least three standardized tests a year, it is estimated. Of these an estimated ten percent are subjected to and are damaged by culturally inappropriate tests. Some researchers have utilized these dubious results to refute the educational validity of the multiracial classroom. Beyond ability grouping is the even more doubtful practice of prediction, using achievement test results. This paper highlights the impact that the testing industry has on curriculum development, especially during the early elementary grades. All of the points referred to above cluster around the issue of test unfairness. There are really two separate issues involved: unfairness in the tests themselves and unfairness in the use of tests and test scores. In this paper, test bias is discussed in three parts: bias due to (1) content factors, (2) bias due to norming, and (3) bias due to the testing situation. Following this, the uses and abuses of tests are discussed along with the political and economic implications of misuse. (Author/JM)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED127382

# STANDARDIZED ACHIEVEMENT TESTING: SOME IMPLICATIONS FOR THE LIVES OF CHILDREN

Robert L. Green*, Julie G. Nyquist**, Robert J. Griffore**

Achievement tests are generally considered to be 20th century phenomena. But looking back into Biblical times, we find that God used a form of testing to select his most able soldiers. He devised a very practical achievement test for General Gideon to administer to his troops in the campaign against the Midianites.

God complained that Gideon's army was so large that the Israelites would think they'd conquered the Midianites themselves, without God's help. To pare down the troops to 300--a small enough number to make a victory look miraculous--God said, "Take the men down to the water, and I will test them for you there. . . Everyone that laps the water with his tongue, as a dog laps, you shall set by himself; likewise everyone that kneels down to drink." Having rejected most of the Army as misfits, God went on with his test. The soldiers he wanted were those who held their heads up, watching for the enemy, while their cupped hands raised the water to their lips. "And the number of

---

*Robert L. Green is Dean of the College of Urban Development and Professor, Educational Psychology, Michigan State University, East Lansing, Michigan. **Julie G. Nyquist and Robert J. Griffore are Ph.D. candidates in the Department of Educational Psychology, Michigan State University, East Lansing, Michigan.

Prepared for National Institute of Education's Test Bias Conference, December 2-5, 1975, Washington, DC.

UD016191

those that lapped, putting their hands to their mouths, was 300 men. . . And the Lord said to Gideon, 'With the 300 men that lapped I will deliver you and give the Midianites into your hand.'" (Judges, Chapter 7)

Down through the years, the Biblical tradition of testing has thrived and prospered, but as we shall discover, the ethic of practicality has been regretably displaced by the doubtful and dangerous morality of expediency. While discrimination continues, the criteria are no longer administered at the waters edge, but rather in the thicket of thorny issues we have come to know as "standardized testing."

The Problem

The scores young children receive on standardized tests can have tremendous impact on their lives, both present and future, guiding or indeed restricting their educational careers, future employment opportunities and their adult lives in general. Because these tests are culture-specific and value-biased, their wide usage for measuring achievement and predicting future success of children has marked political and economic implications, particularly for minorities and those of lower SES origin. By culture-specific and value-biased we mean that standardized tests reflect middle class values and attitudes rather than linguistic, cognitive and cultural experiences common to all groups. One reason for this is that these tests are largely prepared by white, male, middle and upper class Ph.D.'s, so they predictably reflect the lifestyle and world view of the middle and upper class.

Test scores are determined by the shaping effects of one's environment--by the advantages or disadvantages which have comprised a person's life. Minority children score poorly because their cultural experiences do not correspond with those of American middle class society. For example, on the basis of test scores many Spanish-speaking students have been classified as mentally retarded and placed in special education classes. California courts have handed down decisions clearly condemning the use of tests in English to classify children whose native language is not English. Further, children who have reading difficulties or lack conventional middle class experiences have difficulty answering questions in standardized tests. The scores for non-readers may be depressed if a test is given orally, because they have not had a chance to elaborate their structures of knowledge through reading.

Testing is big business in this country. Last year the testing industry reported an income of more than $300 million (Miller, 1974). The industry makes money by selling tests, too often without inserting caveats regarding the limitations and proper uses of the tests. Test constructors must face up to the basic and obvious fact that their efforts do not produce instruments without social implications. There are no ethically or socially neutral tests. And they cannot legitimately avoid the fact that their tests may be used in both fair and unfair ways. Oscar Buros has pointed out that test users are not apparently concerned about the quality of the available

tests as much as they should be. He observes:

> Unfortunately, the rank and file of test users do not appear to be particularly alarmed that so many tests are either severely criticized or described as having no known validity. Although most test users would probably agree that many tests are either worthless or misused, they continue to have the utmost faith in their own particular choice and use of tests regardless of the absence of supporting research or even of the presence of negating research. When I initiated this test reviewing service in 1938, I was confident that frankly critical reviews by competent specialists representing a wide variety of viewpoints would make it unprofitable to publish tests of unknown or questionable validity. Now, 27 years later and five Mental Measurements Yearbooks later, I realize that I was too optimistic. Although many test users undoubtedly are selecting and using tests with greater discrimination because of the MMY's, the publication and use of inadequately validated tests seem to be keeping pace with the population explosion. (Buros, 1965).

While it is basically true that the test user should validate a test on each occasion of its use, the test maker has the primary responsibility for making this clear to the user. Just as the television viewer will only watch the current programming fare, the test user, who may not be sophisticated in test theory or the use of tests, cannot go beyond the information he is given by the test maker. When test makers do not firmly establish the quality of their instruments, they may reveal that their primary concern is with profit-making.

The use of standardized tests in public educational institutions

has become a major national issue. Critics of testing, including minorities and educators, are concerned with test bias, rigid educational tracking, limits to equal opportunity resulting from testing and placement, and the absence of tests that accurately determine whether all children are receiving quality education. Community groups, psychologists and professional organizations have all declared opposition to testing on some related basis. Even state legislatures and the courts have become involved. A strong protest arose from the National Education Association. At their 1973 conference, they endorsed such actions elimination of IQ scores in cumulative records and censure of the testing industry for not responding to the needs of culturally diverse children. They also proposed an immediate moratorium on testing and created a task force to research the subject. The resolution called for an interim cessation of the use of tests for assessing student potential or achievement pending a critical appraisal, review and revision of current testing programs (Bosma, 1973).

Another consistent critic of tests has been the Association of Black Psychologists. At their 1975 meeting last August, that body unanimously resolved not to develop an organizational working relationship with any testing corporation in order to preserve its status as an independent critic of test bias and test abuse. ETS had asked the Association to set up a special committee to work with it to develop cultural-free tests. It was the judgment of the members of the Association, that once the organization estab-

lished a formal agreement with a testing company, the potential was created for co-option because of the pecuniary relationship. Once an employer-employee ambiance is instituted, remaining a credible critic becomes difficult. House critics frequently serve the purposes of the corporation rather than remaining strong, independent voices; and independent voices are needed to prod the lethargic education and testing institutions to alter their actions, in the best interests of minority children. Although no concrete action has emerged from these proposals they indicate that major professional organizations are deeply concerned.

Positive action in restricting the use of tests has been taken by the legislatures of several states. In 1972 the California Assembly enacted provisions instituting a variety of measures designed to ameliorate test abuse, which included prohibiting the inclusion of test scores in cumulative records, establishing bilingual testing and requiring parental consent for placement in classes for the retarded (Miller, 1974). Most importantly, the legislation has directed school districts to use a wider variety of mechanisms and resources to assess potential.

The courts, too, have acted to reform the testing system. One recent decision ruled that tests cannot be used as a basis for classifying students. There have been at least 30 class action suits seeking to force school districts to cease and desist in the inaccurate testing of minority children (Tractenberg, 1974).

Major test corporations, for example the Educational Testing Service, are now acknowledging for the first time the opposition to practices of the testing industry. They have done much in the area of research and some in the area of correcting bias in tests, but they have not demonstrated a clear effort to reduce test misuse. The small efforts that have been made, such as establishing minority committees, are ineffectual. ETS established a Minority Affairs Department, but without budget or power. The effort of such a department is wholly symbolic. This approach lacks the pertinence and strength necessary for coping with the many-faceted issues currently being raised.

Who are the children most likely to be affected by test misuse or abuse? What is the magnitude of such abuse, and what are the issues involved? In this paper, the children of the nation's poor, the "educational underclass" are the people of concern. Black, Puerto Rican, Chicano, native American and low-income white children represent this vast educational underclass. Many American children would come under this heading.

What is the magnitude of the problem of test abuse? It is estimated that more than fifty million American children take at least three standardized tests a year (Goslin, 1967). Of these, an estimated 10 percent, or some five million children are subjected to culturally inappropriate testing methodology and must bear the scars of results that are indelibly imprinted on their future educational and career opportunities.

Further, as we have seen, some researchers have utilized these dubious

test results to refute the educational validity of the multiracial classroom. A case in point is the data employed by Christopher Jencks in his controversial book Inequality. Although to Jencks' credit he cites evidence that school desegregation does promote positive socio-economic gains for minority children, he utilizes testing data to suggest that the achievement curve remains flat and unaffected. What Jencks does not sufficiently discuss are the debilitating results of ability grouping that in many school systems is the result of precisely the test data he cites. Especially in larger school systems such ability grouping is systematically practiced and, of course, it is in the larger systems that the majority of poor and minority children receive their education.

Beyond ability grouping is the even more doubtful practice of prediction, using achievement test results. It is frequently determined by school administrators that a child's educational horizons are limited even though achievement tests often have low predictive validity. This is part of the etiology of what Jonathon Kozol (1967) has called "death at an early age."

It can be well documented that issues related to test use are of concern to all of the constituencies mentioned earlier. Close scrutiny of the literature cited in this paper will provide adequate documentation. This paper will also highlight the impact that the testing industry has on curriculum development, especially during the early elementary grades. . . a period of significant and formative growth and development.

Test Bias

All of the points referred to above cluster around the issue of test unfairness. There are really two separate issues involved; unfairness in the tests themselves and unfairness in the use of tests and test scores. Unfairness in the test is generally referred to as test bias. Thus, a biased achievement test may be defined as a test which does not measure the same dimension(s) of achievement across different groups. (Green, 1975) An example would be a reading comprehension test which measures both comprehension and vocabulary for one group, but only comprehension for a second group, since the latter knew all the terms. Differential prior knowledge of the context could also be a biasing factor in the scores of such a test. (Green & Rondabush, 1975) In this paper, test bias will be discussed in three parts; bias due to (1) content factors, (2) bias due to norming and (3) bias due to the testing situation. Following this, the uses and abuses of tests will be discussed along with the political and economic implications of misuse. Although the two faces of unfairness will be discussed separately they must be considered as interrelated and compounding issues that must be dealt with together.

Bias in the content of a test threatens all forms of test validity, construct, criterion, and content, for the group against which the test is biased. However, since the primary purpose of achievement testing is to measure the student's present status, in relation to a particular subject matter, this discussion will focus on content validity. The central issue in a discussion of content invalidity is: the degree to which certain test items may be more germaine to one group of children (middle SES) than to another (low SES). (Messick & Anderson, 1970) The issues involved are (1) who writes the test items, (2) what groups are the items tried out on, (3) difference in

dialect between the favored middle SES child and the lower SES child, (4) source of irrelevant difficulty of certain test items.

First, who writes the items for standardized achievement tests? Traditionally, the test constructors have been white and middle class. They have been brought up in a particular culture, that of the dominant middle-class, and inevitably this influence is reflected in the items they write. This is a source of bias because there are cultural differences between their products and the lower SES users of the test; differences in "styles of thinking, perceiving and reasoning, and in values and expectations" (Green, 1974). These internal differences can result in what Thomas J. Fitzgibbon (1971) referred to as "cultural blindness." This cultural blindness in reference to low SES children can cause test producers to write items which: (1) may violate a minority child's feelings or (2) are in a context which is not associated with the child's experiences.

Two illustrations might help clarify the problem. The following is a paragraph from the reading comprehension section of the Metropolitan Achievement Test for second graders:

> Once a week we have a very special dinner. Before my mother cooks it she looks at a book. It tells her how to make many good things to eat. Last Sunday, we had a delicious new soup and a chocolate pudding dessert. We all call the book "Mother's best friend" (Wilson & Moulton, 1971).

This passage is as remote as possible from the experiences of the inner-city child, while it is very compatible with the life style of the middle class child. This is cultural bias. Paragraphs like this can scarcely interest the minority child. This is especially important in light of evidence that

indicates that "interest" is as potent a factor in reading comprehension as passage difficulty. (Estes & Vaughn, 1973)

A second example of cultural bias is found in early elementary reading tests which evidence a high frequency of inappropriate vocabulary words. As noted by Weber (1974), some test makers expect second graders to know words like: chimney, ribbon, village, chatting, sapling, baggage, and harvest. While breadth of vocabulary is an aspect of reading skill in the higher grades and in high school, he says "in testing mechanical skills of beginning reading, it penalizes the child with a small hearing vocabulary" and provides a source of cultural bias in the tests against children from the inner city and from homes where a foreign language is spoken. The presence of these rural middle-class words is evidence of bias due to the orientation of the item writer.

This form of bias is, however, correctable. Trained professionals from minority groups must be included more extensively in the production of test items. By trained professionals, we do not mean merely educated and interested minority persons, but those who are expert in test construction and can thus make a real impact in elimination of item writer bias. Since the identity of the test writer is determined by test publishers, this is one area where they can and should be held accountable to minority children.

The second issue in relation to content bias involves the item "try out" conducted during test standardization. Standardized achievement test items are generally "tried out" on a particular sample to detect inappropriate items. Green (1971) conducted research to determine if the choice of

a particular try-out group over other groups would bias achievement tests. One of his conclusions was that "If a biased test is one which contains a substantial proportion of items which would not have been chosen if a different try-out group had been used, then most tests are biased for most children. However, by this criterion of bias most standardized tests are more biased against minority children than against white middle-class children." What this means is that minority children may be required to answer proportionally more questions which are inappropriate for them. As with the problem of who writes test items, the solution to this bias producing problem is within the control of publishers. The publishers should therefore be held responsible for its solution.

Green offers one solution to this problem which includes a suggested solution for the first problem as well: Have separate groups of item writers and editors from all major ethnic and cultural groups each produce a separate trial version of the achievement test. Then try out all materials on each sub-group separately. Finally select items from all versions and edit them to best serve the interest of all groups (Green, 1974). In his estimation, this procedure will lead to tests with the least amount of these forms of bias.

A third source of possible content invalidity is bias due to dialect or language differences between the majority group and minority students: poor Whites and Blacks, Chicanos, and Native Americans. Several researchers have studied dialect differences and test performance and found language differences to be a source of bias (Labov, 1969; Palomares, 1971; McDiarmid, 1971; Cervantes, 1974.) Although this form of bias can be demonstrated in relation to varying areas of achievement, it is of primary interest in the areas of early

elementary reading and reading readiness tests. In reading tests one study of black fourth graders, using Gray's Oral Reading Test, found that when the test was scored according to the regular key, 46% of the errors made by the total group could be attributed to dialect differences. (Hunt, 1972) Many decisions affecting the education and lives of poor children are made on this type of misleading information. To paraphrase William Labov, the chief difficulty is not the dialect differences themselves but the ignoring of those differences. We might add that even more important than the ignoring of differences in dialect, is the making of decisions about children's lives based on this ignorance.

The case of reading readiness tests is even more serious. "On the basis of a single test, which purports to confirm that most minority children are less ready to begin formal reading instruction, most of the minority children are placed in slow groups or not allowed to begin formal reading instruction at all. Thereupon they do not achieve as much as their 'more ready' contemporaries--which is usually taken as proof that they indeed were not as ready to begin formal reading instruction." (Weber, 1974) Use of these tests has tremendous social, economic and political implications for minority children. Under the guise of placing children in "the most appropriate" group for instruction, minority children, already disadvantaged, can be put at even a greater disadvantage through a retarded educational program. Further, in schools where all or most of the children are disadvantaged, the entire curriculum can be thus weakened.

The fourth issue concerns a bias in certain items on achievement tests which results in invalidity due to irrelevant difficulty for lower SES children. It has been found that in many reading comprehension tests there are

two types of items which create bias by presenting the lower SES child with irrelevant difficulty. (Tuinman, 1973-74; pyrezak, 1974) First is the item which can be answered without having read the passage, by using middle-class common knowledge. The other is the item which goes beyond the passage by asking a question which requires both information from the passage and middle-class common knowledge to be answered correctly.

The following is an example of a question that can be answered correctly without reading the accompanying paragraph: "At the museum there were 1. paintings, 2. books, 3. dogs." And this is an example of a question that requires information not provided in a paragraph about a boy who comes from a country where everyone speaks Spanish: "Juan might have come from 1. Germany, 2. England, 3. Mexico, 4. France." The paragraph makes no mention of Mexico so the pupil has to know that Mexico is a Spanish-speaking country. (Weber, 1974)

When questions in any achievement test require the use of middle-class common knowledge as well as knowledge of the specific subject matter "being tested" the test is an unfair measure of achievement in that subject matter area for the low SES child. It might be a perfectly fair test of general information but that was not what the test purported to measure.

A second form of bias relating to the individual test is due to inappropriate test norms, especially those called "national norms." When "national norms" are used, a comparison is made between the performance of a pupil or group of pupils and the norm arrived at through the standardization process. The issue here is not the size of the sample used but the adequacy of the selection of the students included.

No norm will fit any community exactly, however, the dangers of misinterpretation are multiplied if the sample does not include students from all ethnic, racial, regional, community type, and income groups. If the sample leaves out or misrepresents some group, the term "national norm" is a misrepresentation. An example of this problem was evidenced in the standardization sample used for the Metropolitan Achievement Test in 1958. It overrepresented middle-class, rural, and Southeastern students.(Hunter & Rogers, 1967) When "biased" norms such as these are used to make decisions concerning poor or inner-city children needless errors to their detriment can result.

The use of "biased" norms is especially detrimental when achievement tests are used as if they were predictors of future ability or as measures of aptitude. Such inferences can only be made when a student is being compared with persons of the same age from the same sociocultural background. (Mercer, 1974) This is a questionable use of achievement tests, but even if it were not, the use of "national norms" would not be appropriate.

Further problems with the use of norms in test interpretation are caused by misunderstandings about the meaning of the term "norm." The norm is merely the average score of a specially selected sample group, but it is often taken to mean a standard of minimum performance. (Nystrand, 1975) This misconception was illustrated by Dyer. (1973)

> In 1971 the education committee of one of the state legislatures came up with an educational accountability bill that read in part as follows: "If the performance of any school district on any test approved by the state board of education - does not equal or exceed the national performance average of such a test for two successive years, said school district shall not receive any further state finan-

> cial assistance - until such time as said school district has achieved such national performance average.

The bill did not pass the legislature but it demonstrates that norms can be misused even when the unit of analysis is a group and not an individual.

There may be occasions when school systems or researchers have a need to make group comparisons for evaluation purposes; however, there should never be a time when it is necessary to compare individual elementary school children to "national norms." This type of comparison provides no information useful in helping the child, while it can harm him. It is too easy for people to interpret scores as absolute and classify children accordingly.

This problem is aggravated even further when scores are reported as age norms, which are more often and more easily mistaken to be standards of minimum performance. Teachers and principals want all of their students to score at or above grade level. When all do not, which is inevitable, it is tempting to judge the child as inadequate. The scores of young children can be affected by many factors not necessarily related to knowledge. Therefore, to compare underprivileged young children to an absolute, like a national norm, is an unfair use of tests. If interchild comparisons need to be made, local norms, where the children have similar backgrounds, are more appropriate.

Finally, although norms can be useful in analyzing group status, the test user must check carefully to be certain that the standardization sample is appropriate. Also, test producers must be certain that the composition of the norm group is clearly delineated.

A third group of variables, which are apart from test content, can also have an irrelevant influence on test scores and therefore affect test validity. These variables involve the testing situation and can be referred to

as atmosphere variables. (Flaugher, 1974) They include such things as speededness, test-wiseness, answer sheet format, item type, examiner characteristics, perceived use of the test results, and achievement motivation.

A widely mentioned atmosphere variable is speededness, or time allowed for testing. The main issue concerning this variable is: Do the time restrictions of standardized tests have a more aversive effect on the educationally "underclassed" than on other children? Logically it seems that this would be true, however, there is only limited evidence to support this contention. One reason for this lack is that the majority of studies in this area have been conducted with test sophisticated college level students. From this type of study three general conclusions have resulted. (Evans & Reilly, 1972):

1. The tests are somewhat more speeded for predominantly low SES groups.
2. Reducing the amount of speededness produces higher scores for both groups.
3. Reducing speededness is not significantly more beneficial to the lower SES groups.

These conclusions are not surprising in light of the test sophistication of the subjects being studied. By the time students are seniors in college the differences in working speed and test taking ability between students depends very little on their background. All of these students, regardless of background, have had to demonstrate their test taking skills innumerable times to get that far. Those who were handicapped the most are no longer in school. The problem of subject sophistication was admitted by Evans and Reilly (1974) when discussing their own and other relevant studies.

A conclusion contrary to those stated above has been cited in several research studies actually done with naive subjects. The finding was that for a naive test taker significant score improvements are made under unspeeded conditions over speeded conditions. (Knapp, 1963; Moreton & Butcher, 1963) Since all young educationally "underclassed" children may be considered especially test naive, the time limits of early achievement tests could present an important source of test unfairness for these less privileged children.

A second important atmosphere varible is test-wiseness. The contention is that middle SES children are more sophisticated in test taking, regardless of content, than the educationally underclassed childern. There is evidence that this difference does exist and that it can make a difference in test performance. There is also evidence that these skills can be taught to school children. (Callenback, 1973; Evans & Pike, 1973)

There are two basic dimensions of test taking ability. The first is general knowhow, which encompasses such strategies as how to pace yourself, how to avoid unnecessary errors, knowing when to guess, and how to pick a correct answer by eliminating incorrect options. (Millman, et.al., 1965) These strategies are much less typically known by the educationally underclassed student than by the educationally advantaged. This technical lacking can handicap his performance on any standardized test, even those of the highest quality. The second part of test taking skill is what is commonly thought of as test-wiseness, the ability to take advantage of irrelevant clues in items to help answer questions without necessarily knowing the content. The inclusion of irrelevant clues in items can be avoided by strict adherence to good item writing principles such as

those recommended by Ebel (1972). However, not all test constructors follow these rules; and, as Buros indicated, many test users are not careful in their test selections. Therefore the bias due to test wiseness exists and is a problem. An illustration of items which are subject to the effect of test wiseness comes from the Iowa Silent Reading subtest on sentence meaning. Following are some examples from the test:

Do all students have the same determination to achieve?

Does public opinion ever disregard degrees of justice?

Do individuals always adjust themselves to their environment?

Are all anti-trust law enforced with facility?

Is all good writing the result of frequent consultation of an outline?

Is a locquacious individual necessarily a bore?

According to Livingston (1975), approximately 40% of the items in this particular test contained "signal" words like: always, never, all and none, that influence reponses of the test-wise student regardless of his general understanding of these sentences. This subtest obviously shows much bias in favor of the more test-wise advantaged student. Although most achievement tests probably do not show this much bias, it is indicative of the possible magnitude of the problem. Two things could be done to help counteract this type of bias. (1) Careful item construction and choice of test. (2) Special instruction for educationally underclassed children to help them develop test taking strategies.

Such atmosphere varibles as examiner characteristics and perceived use of test results have been found in some research to have a detectable effect

on test performance and motivation of minorities and perhaps on majority members too (Sattler, 1970; Katz, 1972; Savage & Bowers, 1972; Epps, 1974). There is some conflicting evidence but the conclusion generally is that an examiner of the same race is most likely to be facilitative, especially with young children.

A final atmosphere variable is motivation. We cannot distinguish the student's effort or attention to the task from his ability to perform it (Labov, 1969). Most minority and poor children tend to be less test motivated than middle class children, which can affect their test performance irrespective of ability or knowledge (Zigler & Butterfield, 1968).

To summarize, test unfairness can result from bias in test content, inappropriate or misinterpreted norms, or from factors in the testing situation itself. This bias is generally against the poor, minority, or inner-city child. What effect does test unfairness have on the lives of these children? To answer this question it is necessary to discuss first the types of decisions made using standardized tests.

There are many decisions made in school systems, some affecting individual children, some affecting large groups of children, the general curricula, or special academic programs. We will consider four types of decisions: (1) Program planning evaluation, (2) Administrative decisions, (3) Diagnostic decisions, and (4) Classification decisions.

Program planning and evaluation involves decisions relating both to the individual child and to programs in general. These types of evaluative decisions can be classified as either formative or summative (Scriven, 1967).

Formative evaluation has the dual function of monitoring the child's progress and identifying weaknesses in the educational program. It allows the educator to make continuous changes in a program. Tests used for formative evaluation are primarily concerned with changes in skills or attitudes of individual children as a result of the school program. Carver (1974) has introduced a useful distinction between two types of test: psychometric and edumetric. The psychometric test is in widest use and serves the purpose of distinguishing levels of achievement between children; whereas the edumetric test is designed to show growth of knowledge within a child over time. The two tests require different types of items. Psychometric tests call for items of uniform middle difficulty because these items promote the greatest score variance between children while edumetric tests demand items which most children would get wrong before instruction and correct after instruction. For formative evaluation, edumetric tests are more appropriate than the traditional normative or psychometric type of test.

The use of edumetric tests in formative evaluation is an appropriate and necessary application of tests. Children benefit when their instructors are able to follow their individual progress and that of the programs designed to promote their education. Formative evaluation facilitates the early detection of a child in difficulty or a faltering program. It must be noted, however, that the vast majority of available standardized tests are of a psychometric nature. A profitable line of research in testing would be the investigation of the use of edumetric tests as an alternative to psychometric tests for some purposes.

Summative evaluation relates to decisions concerning programs at their completion. This form of evaluation provides the means to judge a program's performance and/or that of its participants. Whereas formative tests measure individual improvement on specific objectives, summative tests measure overall achievement on a wide sampling of program objectives. Bradley and Caldwell (1974) suggest using tests which include items that have a fairly wide range of difficulty to accomodate individuals who differ in how much they benefitted from the program and obtain a more accurate estimate of the program's effect on each individual. Although measurement theory indicates that middle difficulty items are technically superior, it is important with young children to include a number of items of fairly low difficulty to give every child an experience of at least partial success. Reliable measurement instruments are important but not as important as the lives of young children which are shaped, at least in part, by their experiences of success or failure in school.

As a part of their summative evaluations, many schools use standardized tests. These can indicate how the performance of a school or school system compares to other schools or school systems in relation to the subject matter covered by the test. If this information is considered necessary, then the administering of a carefully chosen achievement battery could be appropriate. Tyler (1974), however, offers several criticisms of this type of norm-referenced instrument when used for summative evaluation purposes. 1) A given test does not reflect the particular objectives of the local educational program, method or instructional materials. 2) Norm-referenced tests are not composed of reliable samples of the things that children are

being helped to learn in a given grade, but, rather sample exercises on which children of a given grade differ widely. The things that most children are learning in that grade are likely not to be included in the test sample because of the item selection procedures. 3) The typical achievement test includes too small a sample of exercises appropriate for appraising the learning of children who deviate markedly from the average. The tests include so few items at the level where most disadvantaged children are performing that changes in tests scores due to improved capability cannot be distinguished from those due to chance variations in performance. 4) A teaching method or set of instructional materials is commonly designed to improve learning significantly, but not spectacularly. Most standardized achievement tests are designed to furnish scores sufficiently reliable to identify learning increments taking place in a year but not those that might occur in a fraction of a year.

Standardized achievement tests can be used appropriately in summative evaluation if certain cautions indicated by Tyler's criticisms are heeded. There is a vast array of tests available; careful choice is necessary to assure a good content fit to the local curriculum. Results should be interpreted only for groups of children, not individuals, especially when the children are the educationally underclassed. Finally, these tests should be used cautiously in evaluating programs of less than one year duration.

Administrative decisions comprise a second class of decisions made with the use of standardized tests. These management decisions include such things as space allocation and budgeting for new programs. Decisions are made at many administrative levels and may affect one classroom or an entire

school system. Often the information needed to make management decisions can be obtained without testing the individual children. If, however, a standardized test is used, it should be one that will provide the specific information needed at the administrative level where the decision is being made rather than one that provides a more general kind of information.

A third type of decision made with the help of test information is the diagnostic decision. Test information should, of course, not be the only information used in making any diagnostic judgment. Further, the major aim of diagnostic testing ought not to be program placement, but determining what specifc type of instruction or treatment a child needs. Although this is a legitimate use of tests, it is doubtful that traditional norm-referenced achievement tests are useful for this purpose. These tests can provide information about how a child's performance compares to that of a particular norm group and indicate a comparative weakness in the subject area. However, they are not designed to indicate what type of help a child needs; they are not prescriptive. If decisions of this kind are made using normative achievement tests, that is a misuse. To be of diagnostic value in the classroom, a test would need to be designed to prescribe instruction for particular areas of weakness within the subject matter for each individual student. Diagnostic testing could be considered a special case of formative evaluation where the decision is being made about a child within a more general program. For this reason it again seems that tests of an edumetric nature would be most appropriate.

A fourth set of decisions made with the use of standardized achievement tests is classification decisions, encompassing screening, ability grouping, either within a grade or class, and assignments to special programs. These uses of test data are generally lacking in legitimacy. Normative data are used to make judgments about the abilities of individual children. Although all types of classification decisions have implications for "educationally underclassed" children, ability grouping, because it generally has the strongest negative impact, will be discussed in the section which follows.

In summary, the appropriateness of using standardized achievement tests for decision-making depends upon the type of test used and the kind of decision being made.

Implications of Standardized Testing

Until recently, test writers have been unconcerned about the educational, economic, social and political ramifications of standardized tests. Test writers have typically argued that they are neutral. . .that it is their job to develop instruments that will properly assess children, but that it is up to educators to be concerned about the implications of the use of these instruments. But the impact of assessment, particularly standardized testing is substantial in the lives of elementary school children. Since satisfactory educational experiences early in a child's schooling are prerequisite to later educational success, we will consider first some educational implications of standardized testing.

Educational Implications

There are valid uses for standardized tests. They can be useful in identifying weaknesses in a child's educational background which call for differential instructional treatments. However, tests have been used primarily for purposes of predicting individual's future performance levels. The dubious validity of many available instruments and their biased content jointly work against low SES and minority children. Poor test performance in the early grades is often associated with "low ability" on the part of the child. Consequently, teachers will sometimes reduce their teaching effort, assuming that the child cannot profit from quality education. Thus, many low SES and minority children often have their educational opportunities limited rather than receive the additional help they need to succeed.

The use of standardized tests as a predictor of success can have strong implications for low income groups: the "educational underclass." These are the children that are most often placed in low ability groups. Ability grouping or

tracking is most prevalent in the larger school districts where there is the heaviest concentration of minorities.

Ability grouping began in the United States after World War I. The Army Alpha and Beta exams had focused the attention of educators on the "differences" between people. Added to this was the proliferation of "ability" tests, which allowed the educators to make discriminations between ability levels. The original expectation was that ability grouping would facilitate improved achievement for all children. However, as a result of evidence from numerous research projects during the years 1920-1935, which failed to show that homogeneously grouped students achieved any better than those heterogeneously grouped, ability grouping fell out of favor.

Grouping was practiced minimally between 1935 and 1950 but became popular again in the 1950's after Sputnik promoted an interest in enrichment programs for "bright" children. Although there is no more evidence now than forty years ago to support the use of homogeneous grouping by ability, it unfortunately still enjoys popularity among teachers and school systems.

In 1970 Findley and Bryon carried out an extensive review of the research and literature concerning ability grouping and came to these conclusions:

1. Ability grouping is widely practiced in American school systems. Seventy-seven percent of the school systems surveyed do some ability grouping with 55% generally grouping children. Over 80% of the systems that group by ability use standardized tests either as the sole criterion or along with other criteria.
2. Ability grouping is especially characteristic of larger school systems.

3. Socioeconomic and social class differences are increased by streaming; reduced by non-streaming.
4. Some studies offer positive evidence of effectiveness in high-achieving groups; studies provide almost uniformly unfavorable evidence for promoting scholastic achievement in average or low-achieving groups.
5. The effect of ability grouping on the affective development of children is to reinforce favorable self-concepts of those assigned to high achievement groups, but also to strengthen unfavorable self-concepts in those assigned to low achievement groups. Low self-concept operates against motivation for scholastic achievement in all students, but especially among those from lower socioeconomic backgrounds and minority groups.
6. Assignment to low achievement groups carries a stigma that is generally more debilitating than relatively poor achievement in heterogenous groups.
7. Low-achievement groups tend to receive less intellectual stimulation than do high-achievement groups. (Findley and Bryon, 1970, I)

A special task force directed by Findley also found that:

1. Tests were widely approved by teachers and administrators.
2. Virtually all ability grouping plans depend on tests of aptitude or achievement.
3. The effect of grouping procedures is generally to put low-achievers of all sorts together and deprive them of the stimulation of middle-class children as learning models and helpers.

4. Testing is usually in English which is a second language for many minority children (Findley & Bryan, 1970, IV).

The courts, too, are taking a closer look at tracking and finding such practices unpalatible. In Hobson vs. Hanson (1967), Judge J. Skelly Wright banned tracking in the Washington, DC school system (see section on courts and testing). In his decision he focused on tracking and argued that "when a student is placed in a lower track, in a very real sense his future is being decided for him. The kind of education he gets there shapes his future. . . not only in school, but in society in general. Certainly when the school system undertakes this responsibility it incurs the obligation of living up to its promise to the student that placement in the lower track will not simply be shutting off from the mainstream of education but rather will be an effective mechanism for bringing the student up to his true potential."

School systems have not met this responsibility. It has been documented that a locking-in process occurs once a child is placed in a track. The child often graduates in the lower track or drops out of school entirely. At whatever grade level their education terminates, it often terminates in the track in which they are placed early in their lives (Burry vs. Benton Harbor).

Classification systems based on standardized tests have systematically labeled a disproportionately large number of minority group children as intellectually subnormal and a disproportionately small number of these children as gifted (Findley, 1974). The tests are based on a model which "institutionalizes the culture of the Anglo-Americans as the single monocultural frame of reference for 'normal'" (Hobbs, 1975).

The problem is perpetuated by the attitudes of teachers toward test scores. When teachers perceive children as slow learners, they may assume that

present status must define the future as well. And teachers are amenable to the basic notion of ability grouping. A National Education Association poll among teachers taken in the early sixties reflects this fact: almost 60% of all elementary school teachers favored ability grouping done on the basis of IQ or achievement test scores. In a later poll, 1968, teachers were asked which type of pupils they would prefer teaching, in reference to ability levels. The results among elementary school teachers showed: 18% preferring high ability children; 45% average; 21% mixed ability; 11% with no preference; and only 4% preferring "low" ability children. (NEA, 1968)

If this accurately reflects teacher attitudes it suggests implications for economically underprivileged children beyond the issue of ability grouping. Many standardized tests tend to be content-biased against minority children at least to some degree. Their background also prevents their doing well on these tests. Therefore, a disproportionately large number of minority children receive low scores on standardized tests. Perhaps one-third or possibly one-half of all underprivileged children might be considered by their teachers to fall in a low ability category, and the teachers' attitudes toward these children may have a pernicious effect on their educational achievement. Looked at another way, the "high" ability child has a one in five chance of getting a teacher who prefers teaching him, the "average" child a 50-50 chance, while the "low" ability child get only one chance in 25 of being a preferred student. These attitudes among teachers reflect the dearth of pluralistic teacher training experiences. Ability grouping further promotes an atmosphere of intellectual snobbery from which emerge teachers who desire ability grouping with "someone else" teaching the lower echelon students.

When achievement tests are used as a criterion for ability grouping, it is easily understood why groups concerned with the welfare of children from lower socioeconomic backgrounds call for a complete halt to all standardized

testing.

There are two apparent and promising alternatives to ability grouping on a single dimension. Aptitude X Treatment Interaction techniques are the first and mastery learning is the vehicle for the second.

Aptitude X Treatment Interaction research has suggested that it might be possible to identify traits in individuals with which to optimally match instructional techniques (Cronbach, 1975). While the method might result in complications involving several unintended and unanticipated higher-order interactions, if it is successful, it will avoid the ordering of students according to levels at a single trait, such as reading readiness. Since ATI methodology can be conceptualized as multivariate, it is possible to consider several variables and feedback devices to ensure that various instructional techniques are matched to the students' traits. Further, it would be quite possible and desirable to use a "compensatory model" in bringing all students' skills to high levels of achievement rather than assuming that present status must be destiny.

The other general alternative to undimensional tracking would be some variety of mastery learning approach, perhaps patterned after Bloom's model (Bloom, 1974). The original instruction is presented to a group as a whole without dividing on the basis of any test (thus offering an immediate difference from a system based on tracking). The real advantages in mastery learning come from the variety of teaching techniques used as corrective measures for students who do not attain mastery as a result of the instruction in the large group. The corrective procedures follow feedback from criterion referenced tests and are designed to be matched precisely with the individual student's ability to understand material presented in various ways. For example, while some students learn best by reading either standard prose or programmed material, others may learn better from group discussion at a particular time during their development.

In either the ATI method or the mastery approach, the meaning of the

term diagnosis is different from that commonly associated with the process inherent in assignment according to reading readiness, which only classifies students into one specific level on a single dimension, and where the levels are unfortunately assumed to be hierarchical.

Teacher expectations function as self-fulfilling prophecies. Jere E. Brophy and Thomas L. Good (1970) state that,

> "The teachers demanded better performance from those children for whom they had higher expectations and were more likely to praise such performance when it was elicited. In contrast, they were more likely to accept poor performance from students for whom they held low expectations and were less likely to praise good performance from these students when it occurred, even though it occurred less frequently."

Robert Rosenthal (1973) calls this phenomenon the "Pygmalion effect." Rosenthal has found that students live up, (or down), to their teachers expectations of them. In one study he selected an elementary school in a lower-class neighborhood and administered a non-verbal IQ test at the beginning of the year. The test was disguised as one that would predict "intellectual blooming." There were 18 classrooms in the school, three at each of the six grade levels. The three rooms for each grade consisted of children with above-average ability, average ability and below-average ability.

After the test, he randomly chose 20 percent of the children in each room and labeled them "intellectual bloomers." Each teacher was given the names of those children who could be expected to show remarkable gains during the coming year on the basis of their test scores.

The children were tested again eight months later and it was discovered that those whose teachers had been led to expect "blooming" showed an excess in overall IQ gain of four points over the IQ gain of the control children. It made no difference whether the child was in a high-ability or low-ability classroom. The teachers' expectations benefitted children at all levels.

In another experiment boys and girls, age seven to 14 were tested on their ability to swim. Half of the instructors were led to think that they were dealing with a "high-potential" group, and their students became better swimmers by the end of their two-week camping period than the regular group (Burnham, 1968). Generally, then, a marked improvement can be predicted in students' performance when they are simply labeled "potential bloomers."

Rosenthal has developed a four-factor theory to explain the Pygmalion Effect. He believes people who have been led to expect good things from their students, children, or clients appear to:

--create a warmer social-emotional mood around their "special students" (climate);

--give more feedback to these students about their performance (feedback);

--teach more material and more difficult material to their special students (input); and

--give their special students more opportunities to respond and question (output).

In summary, tests are used as a barrier to full and equal educational opportunity. A low test score on standardized tests can be and often is used as an excuse for a watered-down curriculum from which a student cannot escape. The long-range educational ramifications are that the child can rarely catch up and is unqualified for advanced schooling or for many jobs.

Social, Political, Legal and Economic Issues

Adult life experiences are not ability grouped. A child must learn to work with a wide range of people and to deal with a variety of challenges in our complex society. A student who is placed into special classes may never learn to cope with the realities of life. Most importantly, his self-concept may be adversely affected. The child's status in school and society is determined by his test scores. The image the child develops of himself influences his motivation and aspirations. A positive self-concept helps a child perform satisfactorily on a test; a negative self-concept helps produce low scores. A test situation can be so frustrating to a child that it reinforces negative feelings the student already has toward the educational system.

There is some evidence that the child who does poorly on an achievement test may be so discouraged by his failure that his performance on future tests may be hampered (Bridgeman, 1974). Each new failure experience builds on previous ones in leading a child to view himself as a failure. When a child's academic self-concept declines, his performance in a competitive test situation suffers (Morse, 1963).

Self-concept is related to mental and emotional health. Severe emotional disturbances may result from an individual's sense of failure. Consistent success in school over a number of years "immunizes a child against mental illness;" conversely, those who are consistently unsuccessful "are vulnerable to emotional illness" (Kirkland, 1972).

Failure also affects a student's opinion of his peers and the attitudes his peers have about him. A child who is perceived as unsuccessful

develops difficulty in relating to his peers because he feels inferior to them. Further, his peers come to view him less favorably.

A child with low self-esteem has low aspirations. Failure deflates a child's desire to learn and to strive for high levels of achievement. As individuals meet with success in certain tasks, their goals and aspirations rise in accordance with increased confidence in their abilities to achieve their goals. Thus, a child who does poorly in elementary and high school will not have the motivation to continue his studies at the college level (Kirkland, 1972). This is important, since highly educated individuals achieve higher occupational status than those who are denied equal educational opportunities. People who have not been well educated and have failed in school discover that their chances of finding meaningful employment are slim. In today's society it is impossible even to enter a trade without a high school diploma, so the educationally handicapped individual has no opportunity for upward mobility. Unprepared for the labor market, these individuals have no alternative but to join the growing unemployment and welfare lines. They are forced to remain on the outside of the system. The strength of the relationship between access to excellence in educational experiences and occupational status has not escaped the attention of minority and low SES groups in recent years. And, as Chuck Stone (1975) observes, the courts have demonstrated that they can be effective agents in exposing and remedying the misuse of tests and test scores. For example, one of the most notable sections of the opinion handed down in the Larry P. v. Wilson Riles case (1972) was the contention that "irreparable harm" ensues when a child is placed even temporarily in a class for the mentally retarded on the basis of test scores. For whether or not the child is correctly

placed, he will inevitably feel humiliated, and other students are likely to subject him to some degree of ridicule.

Aside from this basic psychological effect on the child, two other educational issues have found their way into the courts: the use of tests as a basis for tracking and the use of achievement tests as predictors of success. On the former issue, in the 1967 Hobson v. Hansen case (1967), District Court Judge J. Skelly Wright ordered the tracking system in Washington DC abolished because the basis of the assignment was the use of culturally biased tests.

The 1970 Diana v. the California State Board of Education decision handed down by District Court Judge Robert F. Peckham ordered that all children whose primary language is other than English must be tested in that language in addition to in English. The decision was founded on the fact that Spanish-speaking children tested in Spanish may score as much as 30 points higher than when tested in English.

In the Moses v. Washington Parish School Board et al case (1971), the practice of using the Primary Mental Abilities Test and the Ginn Reading Readiness Test to form homogenous groups for reading instruction was challenged. The court ruled that the practice violates the Fourteenth Amendment rights of blacks to be treated equally with whites. And research shows that grouping offers no advantages to children of low ability (Heathers, 1969).

The Larry P. v. Wilson Riles case referred to above resulted in a decision that "No black student may be placed in an EMR class on the basis of criteria which rely primarily on the results of IQ tests as they are currently administered."

Finally, the McNeal v. Tate County School District case (1975) in Mississippi resulted in a decision that ability grouping which resulted in racially segregated classrooms could not be used by a previously segregated school district until it had been integrated without tracking long enough so that children were not assigned to slower groups because of educational disparities brought on by prior segregation.

The second general educational issue recently attended to in the courts is the use of achievement tests as predictive instruments. In connection with employment practices, the Griggs et al v. Duke Power Company case (1971), in Prospect, North Carolina, the ruling made invalid the use of ability tests which resulted in rejection from employment of a disproportionage number of blacks. The relationship between the criterion test's content and the requirements of the job must be firmly established as well.

Equally pertinent to education is the Baker et al v. Columbus Municipal Separate School District, et al case, which considered the validity of test scores as related to job performance. The case established that the use of the National Teacher Examination created a racial classification. Most important of all, the ruling pointed out that it is not yet determined whether a relationship exists between academic preparation, performance on the NTE, and good teaching. The essence of this point is that achievement tests are not justifiably used unless they are known to have good predictive validity.

This brings us to the point made by Thomas Fitzgibbon concerning the political use of tests: "There are no such things as test results which cannot be used for political purposes" (Fitzgibbon, 1973). Our contention is that test results of dubious merit are all too often used politically to the

detriment of poor children. Those in opposition to desegregation and busing are now using test results of this kind to demonstrate that desegregation does not work. The existence of biased tests makes it easier for these people to deprive children of their right to be fully educated.

The purpose of this entire paper has been to emphasize that, as Marian Wright Edelman, President of the Children's Defense League, maintains, children have definite rights. And they are gaining a substantial group of advocates in their favor. Jane Mercer (1974) has discussed five rights of children which are frequently violated by current testing practices. They are as follows:

1. The right to be assessed as a multidimensional human being. For example, in California it had been standard procedure to classify children as mentally retarded on the basis of measured IQ. Of course, this retardation was school-specific rather than equally influential on all behavior. Recently a new California law has been enacted which requires assessment of a child's non-school abilities as well as his school abilities before a child can be considered retarded. In New York, though, to cite a contrary practice, it would be accurate to say that the overriding concern is the production of generation upon generation of "exam passers," since the Regents Exams are the criteria which determine whether students qualify for college. Goals of good adjustment and personal happiness seem to retain the customary back seat so long associated with second class status.

2. The right to be fully educated. For many children assessment has been a direct route to educational mediocrity.
3. The right to be free of stigmatizing labels, such as learning disabled, mentally retarded, speech impaired, etc. Bradley and Caldwell document the unanticipated and undesirable effects of these labels (Bradley and Caldwell, 1974).
4. The right to ethnic identity and respect. Standardized tests emphasize competencies valued in Anglo societies. They do not deal with skills valued in other cultures.
5. The right to be evaluated within a culturally appropriate normative framework. Comparing minority children against majority norms is misleading evidence on their culturally relevant abilities.

In considering these basic rights of children, it becomes clear that with respect to testing, the most important issues are those directly associated with the rights of those children who are alleged to benefit from current testing practices, but who are likely to require strong protection from these same practices. Social issues, political issues, as well as economic and legal issues, are only issues because there is a difference of opinion concerning the degree of and equality of benefits accorded to children of varying backgrounds. It is our position that neither adequate degree nor clear equality of benefits have been demonstrated for low SES and minority children.

Nor is it true that parents' rights are fairly observed in the course of current testing practices. The poor or minority parent has little personal or legal authority. These parents seldom are allowed to see the tests

administered to their children. The social worker can see the tests, as can the teacher who is often poorly informed in test interpretation. Lawyers, curriculum personnel, and others may also gain access to this information. But even if the parent is able to see the test scores, all pertinent information regarding the test is sacrosanct.

Poor parents have rights and must be respected as individuals. The poor and disenfranchised have the same right to dignity and respect as the wealthy. They also have the right to see that their children are respected instead of exploited. And the testing industry should assume ample responsibility in seeing that this right is realized.

Some Proposals

On the basis of the evidence cited in the development of our position, we believe that some suggestions are in order, all of which would be quite likely to correct many of the problems described above.

1. We agree with Hobbs (1975) that it would be appropriate to establish a National Bureau of Standards for Educational and Psychological Tests and Testing. Such a review agency is necessary in a society which has come to rely on testing as the primary standard for selection, classification and placement decisions in all aspects of citizens' lives.
2. It may well be a wise and ultimately profitable practice for NIE to allocate funds for conducting independent validation and reliability studies on the widely used educational psychological instruments.

3. There should be full involvement of minority professionals in developing, revising, and reviewing standardized achievement tests.
4. Corporations which produce standardized tests should take an active approach toward limiting test misuse. The assumption of a neutral position is not realistic.
5. All tracking should be eliminated because it results in educational and social class division.
6. Test companies should join with school systems and educators as well as measurement specialists to seek ways in which instruments, edumetric among others, can be designed to enhance the educational status of all children through improvement of instruction.
7. The use of national norms on achievement tests for individual decision making and diagnosis should cease.
8. Testing companies should adopt the policy that their purpose is the full development of children. Their concerns should incorporate personal, social and intellectual development.
9. Educators in the public school systems and universities must acknowledge the social, political and economic ramifications of testing. Tests in themselves do make a difference in children's lives.
10. A major portion of the economic resources of the rich and thriving testing companies should be earmarked for research and

development in the area of test unfairness and bias in achievement tests. Indeed, if these recommendations are not carefully considered, it is quite probable that other independent investigations will occur. Chuck Stone (1975) has suggested that a Congressional investigation is in order.

11. It is time to expect and indeed demand that test producers and test users should be primarily concerned with what standardized achievement tests can and cannot do. We need a new emphasis on "truth in testing."

BIBLIOGRAPHY


Baker, et. al. v Columbus. Municipal Separate School District, et. al., 329 F. Supp. at 721.

Bloom, Benjamin S. An Introduction to Mastery Learning. In J.H. Block (ed.) Schools, Society and Mastery Learning, New York: Holt, Rinehart and Winston, 1974.

Bosma, Boyd. The NEA Testing Moratorium. Journal of School Psychology, Vol. 11, No. 4, 1973, pp. 304-306.

Bradley, Robert H. and Caldwell, Bettye H. Issues and Procedures in Testing Young Children. TM Report No. 37, Eric Clearinghouse on Tests, Measurement and Evaluation, Princeton, New Jersey, December, 1974.

Bridgeman, Brent. Effects of Test Score Feedback and Immediately Subsequent Test Performance. Journal of Educational Psychology, Vol. 1, No. 1, 1974, pp. 62-66.

Brophy, Jere E. and Good, Thomas L. Teachers' Communication of Differential Expectations for Children's Classroom Performance: Some Behavioral Data. Journal of Educational Psychology, Vol. 61, No. 5, 1970, pp. 365-374.

Burnham, J. Effects of Experimenters' Expectancies on Children's Abilities to Learn to Swim. Unpublished Masters Thesis, Purdue University, 1968.

Buros, Oscar K. (ed.) 6th Mental Measurements Yearbook. Highland Park: Gryphon Press, 1965.

Burry v School District of the City of Benton Harbor, Michigan, February, 1970. See testimony of Robert L. Green, expert educational witness.

Callenbach, Carl. The Effects of Instruction and Practice in Content-Independent Test-taking Techniques upon the Standardized Reading Test Scores of Selected Second Grade Students. Journal of Educational Measurement, Vol. 10, No. 1, 1973, pp. 25-30.

Carver, Ronald P. Two Dimensions of Tests: Psychometric and Edumetric, American Psychologist, 1974, 29 (7), pp. 512-518.

Cervantes, Robert A. Problems and Alternatives in Testing Mexican American Students. National Institute of Education (HEW), Washington, D.C., April, 1974, ERIC, ED 073 951.

Cronbach, Lee J. Beyond the Two Disciplines of Scientific Psychology. American Psychologist, 30 (2), pp. 116-127.

Diana v California State Board of Education. Civ. No. 71-2897 (E.D. La., order issued April 24, 1973).

Dyer, Henry S. Recycling the Problems in Testing. In Proceedings of the 1972 Invitational Conference on Testing Problems, Princeton, New Jersey: ETS, 1973.

Ebel, Robert L. Essentials of Educational Measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Epps, Edgar A. Situational Effects in Testing. In Miller, L.P. (ed.) The Testing of Black Students: A Symposium. Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Estes, Thomas H. and Vaughan, Joseph L. Reading Interest and Comprehension: Implications. Reading Teacher, March, 1973, pp. 149-153.

Evans, Franklin R. and Pike, L.W. The Effects of Instruction for Three Item Formats. Journal of Educational Measurement, Vol. 10, 1973, pp. 257-272.

Evans, Franklin R. and Reilly, Richard R. A Study of Speededness as a Source of Test Bias. Journal of Educational Measurement, Vol. 9, No. 2, 1972, pp. 123-131.

Evans, Franklin R. and Reilly, Richard R. The Affects of Test Time Limits on Performance of Culturally Defined Groups. ERIC, ED 102 276, September, 1974.

Findley, Warren G. Grouping for Instruction. In Miller, L.P. (ed.) Testing of Black Students: A Symposium, Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Findley, Warren G. and Bryon, Miriam M. Ability Grouping 1970: Part I Common Practices in the Use of Tests for Grouping Students in Public Schools. ERIC, ED 048 381, 1970.

Findley, Warren G. and Bryon, Miriam M. Ability Grouping 1970: Part IV Conclusions and Recommendations. ERIC, ED 048 384, 1970.

Fitzgibbon, Thomas J. Political Uses of Educational Test Results. Address presented at 1973 APGA Convention held in San Diego, California, February 12, 1973.

Flaugher, Ronald L. Some Points of Confusion in Discussing the Testing of Black Students. In Miller, L.P. (ed.) Testing of Black Students: A Symposium, Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Goslin, David A. Teaching and Testing. New York: Russel Sage Foundation, 1967.

Green, Donald R. Racial and Ethnic Bias in Achievement Tests and What to Do About It. Princeton, New Jersey: ERIC, ED 084 285, 1974.

Green, Donald R. Racial and Ethnic Bias in Test Construction. Final Report, 1971, ERIC, ED 056 090.

Green, Donald R. What Does it Mean to Say a Test is Biased? Presented at AERA, Washington, D.C., March 31, 1975.

Green, Donald R. and Draper, John F. Exploratory Studies of Bias in Achievement Tests. Paper presented at the Meeting of APA, Honolulu, 1972.

Green, Donald R. and Rondabush, Gordon E. An Investigation of Bias in a Criterion-referenced Reading Test. Paper presented at the meeting of AERA, Washington, D.C., 1975.

Griggs, et. al., v Duke Power Company, 401 U.S. 424, 1971.

Heathers, Glen. Grouping. In Encyclopedia of Educational, 4th ed., New York: Macmillan, 1969, pp. 559-570.

Hobbs, Nicholas. The Futures of Children. San Francisco: Jossey-Bass, Publishers, 1975.

Hobson v Hansen, 269 F. Supp., 401, 1967.

Hunt, Barbara C. Black Dialect and Third and Fourth Grader's Performance on the Gray Oral Reading Test. Reading Research Quarterly, 25:430-437, February, 1972.

Hunter, L.B. and Rogers, F.A. Testing: Politics and Pretense. The Urban Review, 1967, 2 (3) pp. 5-6, 8, 25-26.

Jencks, Christopher. Inequality. New York: Harper & Row, 1973.

Katz, I. Experimental Studies of Negro-White Relationships. In Berkowitz, L. (ed.) Advances in Experimental Social Psychology, Vol. 5, New York: Academic Press, 1970.

Kirkland, Marjorie C. The Effects of Tests on Students and Schools. Review of Educational Research, Vol. 41, No. 4, 1974, pp. 303-350.

Knapp, R.R. The Effects of Time Limits on the Intelligence Test Performance of Mexican and American Subjects. Journal of Educational Psychology, Vol. 33, 1963, pp. 22-30.

Kozol, Jonathan. Death at an Early Age. Boston: Houghton Mifflin, 1967.

Labov, William. A Study of Non-Standard English. Center of Applied Linguistics, Washington, D.C., 1969, ERIC, ED 024 053.

Larry P. v Wilson Riles. 343, F. Supp., 1306, 1972.

Livingston, Howard F. What the Reading Test Doesn't Test -- Reading. Journal of Reading, Vol. 15, No. 6, 1972, pp. 402-410.

McDiarmid, G.L. The Hazards of Testing Indian Children. 1971, ERIC, ED 055 692.

McNeal v Tate County School District. 508 F. 2d 1017, 1975.

Mercer, Jane. A Policy Statement on Assessment Procedures and the Rights of Children. Harvard Educational Review, Vol. 44, No. 1, February, 1974, pp. 125-141.

Mercer, Jane. Sociocultural Factors in Educational Labeling. Paper presented at the NICHD Conference in Niles, Michigan, April 18-20, 1974.

Messick, Samuel and Anderson, Searvia. Educational Testing, Individual Development and Social Responsibility. Princeton, New Jersey: ERIC, ED 047 003, November, 1970.

Miller, L.P. (ed.) Testing of Black Students: A Symposium. Englewood Cliffs, New Jersey: Prentice-Hall, 1974.

Millman, J.; Bishop, C.H. and Ebel, R. An Analysis of Testwiseness. Educational and Psychological Measurement, Vol. 25, 1965, pp. 707-726.

Moreton, C.A. and Butcher, H.J. Are Rural Children Handicapped by the Use of Speeded Tests in Selection Procedures? British Journal of Educational Psychology, Vol. 33, 1963, pp. 22-30.

Morse, R.J. Self Concept of Ability, Significant Others and School Achievement of Eighth-Grade Students: A Comparative Investigation of Negro and Caucasian Students. Unpublished Masters Thesis, Michigan State University, 1963.

Moses v Washington Parish School Board, et. al. F. Supp., 1340, 1971.

NEA Research Division, Teacher Opinion Poll, NEA Journal, February, 1968, p. 53.

Nystrand, Martin. The Politics of Rank Ordering. English Journal, Vol. 64, No. 3, March, 1975, pp. 42-45.

Palomares, U.H. A Critical Analysis of the Research on the Intellectual Evaluation of Mexican-American Children. 1965, ERIC, ED 027 097.

Pyrezak, Fred. Passage-dependence of Multiple-choice Items Designed to Measure the Ability to Identify the Main Idea of a Paragraph: Implications for Validity. Educational and Psychological Measurement, 1974, 34, pp. 343-348.

Rosenthal, Robert. The Pygmalion Effect Lives. Psychology Today, September, 1973, 7 (4), pp. 56-63.

Sattler, J.M. Racial Experiment Effects in Experimentation, Testing, Interviewing and Psychotherapy. Psychological Bulletin, 1970, 73, pp. 137-160.

Savage, J.E., Jr. and Bowers, N.D. Testers Influence on Children's Intellectual Performance, 1972, ERIC, ED 064 329.

Scriven, Michael. The Methodology of Evaluation. In Tyler, Ralph (ed.) Perspectives on Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation, No. 1, Chicago: Rand McNally, 1967.

Stone, Chuck. Let's Abolish I.Q. Tests, SATS (and ETS, Too). The Black Collegian, December, 1975.

Tractenberg, Paul L. and Jacoby, Elaine. Pupil Testing: A Legal View. Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurements and Evaluation, December, 1974.

Tuinman, J.J. Determining the Passage Dependency of Comprehension Questions in Five Major Tests. Reading Research Quarterly, 1973-1974, 9, pp. 206-223.

Tyler, Ralph W. and Wolf, Richard M. (eds.) Crucial Issues in Testing. Berkeley, California: McCutchan Publishing Corporation, 1974.

Weber, George. Uses and Abuses of Standardized Testing in the Schools. Occasional Papers, No. 22, Washington, D.C.: Council for Basic Education, May, 1974.

Williams, Robert L. Black Pride, Academic Relevance and Individual Achievement. The Counseling Psychologist, 1970, 2, pp. 18-22.

Wilson, Susan N. and Moulton, Elizabeth W. Unfair Tests. New York Times, September 18, 1971.

Zigler, E. and Butterfield, E.C. Motivational Aspects of Changes in I.Q. Test Performance of Culturally Deprived Nursery School Children. Child Development, Vol. 39, 1968, pp. 1-14.